Clusters aplicado al análisis de la música y su relación con variaciones macroeconómicas

Resumen

El reto del proyecto es agrupar características musicales y macroeconómicas que permitan ver relaciones entre el tipo de música escuchada, sus características y la situación económica del momento.

Se trabajó con datos de Estados unidos entre 1999 y 2020, relacionados con características musicales de canciones de la plataforma Spotify y macroeconómicas del Banco mundial. Se dividieron en años de recesión y de No recesión para el análisis, de acuerdo con la referencia del Comité de Datación del Ciclo Económico de la National Bureau of Economic Research.

Se obtuvieron resultados sobre las preferencias de géneros musicales en las épocas. Tambien se encontró que la bailabilidad es una de las características diferenciadoras entre épocas. Canciones más bailables son más populares durante No recesion.

Se obtuvieron grupos con características diferentes, tanto económicas como musicales, que pueden generar impacto en la industria musical al servir como insumo para conocer preferencias de la población de acuerdo con el contexto económico.

Se logró una buena aproximación usando los algoritmos K medias y K medioides para las épocas de recesión y no recesión, ya que con ellos se obtuvo la formación de grupos moderadamente compactos y diferenciados.

Introducción

En la era digital, la música se ha convertido en una parte integral de la vida cotidiana, y plataformas como Spotify han transformado la forma en que consumimos y descubrimos música. La plataforma recopila una variedad de información sobre las canciones, incluyendo atributos como el género, la bailabilidad, la energía y la popularidad, así como las emociones que transmiten. Este proyecto se centra en el análisis de estas variables y su vinculación con variables macroeconómicas, con el objetivo de identificar las características musicales en las épocas de recesión y no recesión en Estados Unidos con datos del periodo entre 1999 y 2020. Se busca descubrir patrones y grupos dentro de la música, que se vinculen con cambios económicos, lo que puede proporcionar valiosas percepciones sobre cómo ciertas características musicales se relacionan con el periodo económico del país. La motivación para resolver esta pregunta radica en la posibilidad de aplicar técnicas de aprendizaje no supervisado, específicamente el clustering, para revelar estructuras ocultas en los datos musicales que permitan caracterizaciones relacionadas con la realidad económica de la audiencia. Identificar patrones y grupos de canciones con características comunes puede tener implicaciones significativas para la industria musical, especialmente para los productores y las plataformas de streaming (En este caso Spotify). Respecto a la literatura revisada, este trabajo sería innovador, ya que se encontraron aplicaciones de este tipo de algoritmos en variables de origen musical (Liu& Chao,2021), (Kim, K., Yun, W., & Kim, R ,2015) y (*Poonia* y Malik, 2022), pero no se encontró sobre su vinculación con variables macroeconómicas como las que aquí se presentan. Sin embargo, ya se ha afirmado, que las canciones que escuchan pueden reflejar lo que piensan y sienten las personas sobre la economía (Moreno, R. (2018, Jul 09)). Se encontraron para épocas de recesión y de No recesión 5 clusters. Entre épocas no se encontraron diferencias en las preferencias de los géneros musicales, siempre predominando el Pop, Hip-Hop y Rock, sin embargo, durante la No recesión la música latina predomina de manera similar al Rock.

Aunque se pudieron encontrar insumos interesantes con el enfoque trabajado en el proyecto, en trabajos posteriores se puede buscar afinar los algoritmos y explorar el uso de DBSCAN para la época de recesión con el fin de minimizar el efecto de los datos atípicos presentes en la base de datos considerando que este también presentó un Silhouete score aceptable (0.5944).

Materiales y Métodos

Datos: El conjunto de datos usado proviene de 3 fuentes fusionadas, filtrando información de Estados Unidos para una década(1999-2020).Una es la base de datos de Kaggle sobre música de Spotify (https://www.kaggle.com/code/varunsaikanuri/spotify-data-visualization/comments) "songs_normalize.csv" (Fuente 1), otra informacion se extrajo al conectarse directamente a la API de Spotify (Fuente 2) y finalmente la Fuente 3 fue informacion macroeconómica del banco mundial (https://datos.bancomundial.org/). La base de datos final contaba con 18 variables de interés y 1940 filas. A continuación, se muestra información sobre las variables iniciales al unir las bases de datos, para mayor detalle ver anexo 1.

Variables	Tipo de variable	Clasificación
Duración_ms,anio, popularidad,bailable, energía, volumen, locuacidad,acusticidad, instrumentalidad, vivisidad, valencia emocional, tempo, followers, día_semana	Numérica	Musical
Explicito, modo, clave, genero, artista, cancion	Categórica	Musical
Inflation, unenployment, gdp_per_capita, gross_savings,crecimiento_renta_neta, consumption_expediture	Numérica	Económica.

Se llevó a cabo la eliminación de duplicados (59) y no se encontraron valores faltantes. Todo el procesamiento de los datos y los análisis se realizaron en un Jupyter Notebook usando Python.

Preprocesamiento de datos: Se realizó el análisis exploratorio de las variables que incluyó el análisis de las estadísticas descriptivas, la identificación de datos atípicos mediante boxplot, diagrama de correlación y análisis de frecuencia.

Se encontró presencia de datos atípicos principalmente en variables como popularidad, volumen, instrumentalidad, acusticidad, locuacidad, vivisidad. Se encontró correlación entre características musicales como la valencia emocional y bailabilidad, también entre los géneros que se pueden considerar opuestos como hip-hop y pop, y entre si las variables económicas, en este punto no se visualiza relaciones entre las variables musicales y económicas. Se identificaron variables con alta cardinalidad como artista (835 clases), canción (1879 clases) y genero (59), para el género, se procedió a identificar el género principal como el inicialmente identificado en la base de datos y se cambió el nombre a genero_principal resultando 12 categorías. Se eliminaron las variables que se consideró no aportaban al objetivo del análisis ni al algoritmo: Artista, canción, fecha_lanzamiento y día_semana. Se incluyó una variable que definía si hubo o no recesión en el año. Se obtuvo la nueva base de características seleccionadas:

Variables	Tipo de variable	Clasificación
Duración_ms,anio, popularidad,bailable, energía, volumen, locuacidad, acusticidad, instrumentalidad, vivisidad, valencia emocional, tempo, followers.	Numérica	Musical
Explicito, modo, clave, genero_principal, Recesion	Categórica	Musical
Inflation, unenployment, gdp_per_capita, gross_savings,crecimiento_renta_neta, consumption_expediture	Numérica	Económica.

Las variables categóricas explícito, modo y clave ya venían codificadas por lo cual no fue necesario preprocesarlas, en el caso de genero_principal se utilizó one hot encoding. Se analizó la estadística descriptiva de la base de datos final y se pueden identificar las diferentes escalas de las variables. antes de usar en el modelo de clustering se utilizó escalado, mediante StandarScaler de Scikit-learn para que mayores escalas como por ejemplo duracion_ms (va desde 113,000 hasta 484,146) no influyera más en la formación de los clusters frente a otras variables con escala menor como popularidad 0 a 89 o bailable entre 0.129 y 0.975.

Explícito, modo, clave y genero_principal por su naturaleza no fueron escaladas.

Estadísticas descriptivas

	duracion_ms	anio	Popula-ridad	bailable	energía	volumen	locuacidad	acusticidad	instrumentalidad	valencia emocional	tempo	vivisidad	Followers
count	993	993	993	993	993	993	993	993	993	993	993,00	993	993
mean	218.104	2.014	59,097	0,669	0,712	-5,470	0,100	0,131	0,011	0,514	121,75	0,180	21.757.460.000.000
std	35.532	2,874	25,715	0,130	0,150	1,906	0,092	0,178	0,072	0,214	25,677	0,132	28.359.200.000.000
min	113.000	2.010	0,000	0,180	0,055	-20,514	0,023	0,000	0,000	0,038	64,934	0,022	352.000.000
25%	196.664	2.012	57,000	0,588	0,614	-6,444	0,041	0,016	0,000	0,355	100,58	0,093	2.604.019.000.000
50%	215.064	2.014	69,000	0,677	0,731	-5,242	0,059	0,058	0,000	0,518	122,77	0,125	10.281.850.000.000
75%	234.653	2.017	76,000	0,755	0,825	-4,144	0,117	0,174	0,000	0,674	134,04	0,240	26.663.600.000.000
max	484.146	2.019	89,000	0,964	0,985	-0,740	0,530	0,945	0,901	0,966	205,57	0,853	120.903.900.000.000

	inflation	unemployment	gdp_per_capita	gross_savings	crecimiento_renta_neta	consumption_expenditure
count	993	993	993	993	993	993
mean	1,79	6,24	56.291	18,38	2,79	82,5
std	0,74	2,06	5.278	1,50	1,00	1,3
min	0,12	3,67	48.651	15,17	1,06	81,0
25%	1,46	4,36	51.784	18,13	2,05	81,5
50%	1,81	6,17	55.304	18,67	2,91	82,0
75%	2,13	8,07	60.322	19,45	3,54	83,4
max	3,16	9,63	65.548	19,75	4,31	84,9

El dataframe fue dividido para su analisis independiente, de acuerdo con los años de lanzamiento y su coincidencia con épocas de recesión (191 registros) y bonanza económica (1749 registros) en Estados Unidos. La división se hizo de acuerdo con la referencia del Comité de Datación del Ciclo Económico de la National Bureau of Economic Research NBER (2024), esto permite analizar cómo varía el perfil musical en función de las condiciones económicas. Este enfoque permite aplicar el clustering por separado a cada conjunto de datos y observar posibles diferencias en el comportamiento musical.

Reducción de dimensionalidad: Se utilizó componentes principales (PCA), para retener la mayor variabilidad de los datos, reduciéndolos a la menor cantidad de factores o componentes, mediante grafico de codo se seleccionó el número de componentes adecuado (Plan,1986).

Cluster jerárquico: Inicialmente se aplicó, para obtener una visión general de cómo se agrupan los datos sin especificar el número de clusters de antemano. Se usó la distancia euclidiana y el método Ward para minimizar la varianza y generar clusters más compactos. Con el dendograma, se identificó el número de clusters a usar posteriormente en los algoritmos K-Means y K-medoides. Posteriormente se asignaron las etiquetas de clusters a la base de datos original.

En el caso de DBSCAN, se evaluaron entonces min_samples entre 2 y 10. Usando vecino más cercano para el cálculo de las distancias, se generó un gráfico de codo para determinar el valor optimo de eps y se identificó el punto mediante la función KneeLocator, con estos parámetros se aplicó DBSCAN y calculó el coeficiente de Silhouette como métrica de la calidad de separación de los clusters.

Algoritmos de clustering: Se evaluaron 3 algoritmos para los datos clasificados en las dos épocas.

- K-Means: Agrupación de datos basado en centroides usando distancia euclidiana (Neira,2021), sensible a la asignación inicial del k.
- K_Medoids: Agrupación basada en datos (Medoides) usando distancia euclidiana.Mas robusto ante los datos atípicos.
- DBSCAN: Este método basado en la densidad se utilizó para identificar los clusters de los datos y los puntos de ruido (cluster "-1"). Los parámetros eps y min_samples se determinaron como se explicó anteriormente. Se evaluó por ser especialmente útil en datos con valores atípicos que se consideran ruido por estar alejados de los núcleos y puntos frontera (Quiala Fonseca,2023)

Posteriormente se evaluó el Silhouette score para los 3 algoritmos en recesión y no recesión para obtener el mas adecuado en cada caso de acuerdo con el mayor valor de la métrica.

Visualización: Los resultados se representaron por gráficos de dispersión en 2 dimensiones usando PCA. Posteriormente se analizaron las variables originales en los diferentes grupos.

Resultados y Discusión.

El análisis realizado se aplicó a dos dataframes, resultantes de dividir el dataframe original en época de recesión y no recesión (Bonanza), En ambos casos al aplicar componentes principales se seleccionaron 15 componentes que para el primer caso explicaron el 99% de la varianza y para el segundo el 90%, lo que indica que se conserva la mayor parte de la información dada por este grupo de datos al reducir la dimensionalidad con PCA.

Análisis de dendogramas

época	Rango distancia Euclidiana de corte	Clusters óptimo
Recesión	1000-2000	5
No recesión	5000-10000	5

En ambos casos el número de clusters óptimo está entre 4 y 5, se seleccionaron 5 clusters considerando que sería un valor óptimo para tener grupos suficientemente diferenciados para correr los algoritmos K-Means y K-medoids.

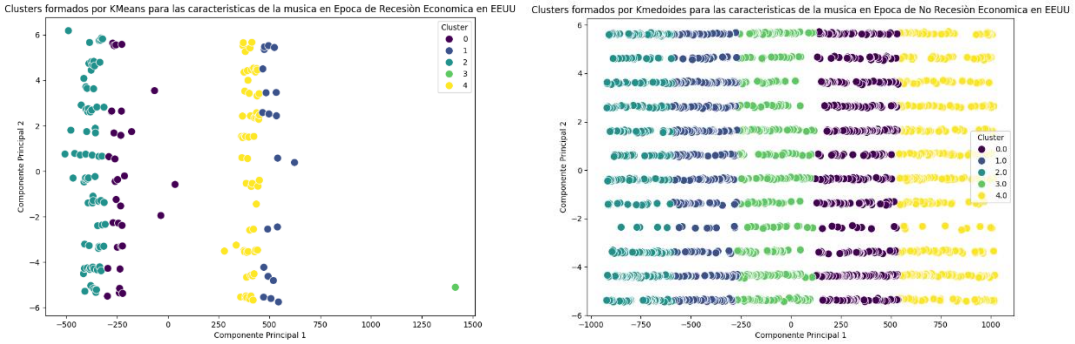
En el caso de DBSCAN para épocas de Recesión la mejor configuración en el rango de iteraciones evaluado (2-10), fue min_samples = 4 y eps = 13.8837. Este valor de min_samples y eps proporciona 4 clusters (coherente con el clustering jerárquico) y un Silhouette Score razonablemente alto (0.5038). Aunque hay algunos outliers (15 puntos), no son excesivos comparado con el tamaño total del dataset. Para No recesión la configuración fue min_samples=4 y eps=10.3194 Aunque el número de clusters es un poco mayor que el esperado, el Silhouette Score es aceptable para DBSCAN (0.456), y se logra una buena reducción del número de clusters en comparación con los primeros resultados de la iteracion (115 clusters).

Al evaluar el Silhouette score para los tres algoritmos se encontró un valor más alto para K-means en el caso de las épocas de Recesión y para K-medoides en las de No recesión. Así se escogieron los algoritmos que para cada época logran una mejor diferenciación de clusters(más separados entre grupos y compactos entre los datos que los conforman):

Épocas	**Algoritmo de clustering**	**Silhouette score**
Recesión	**K-means**	**0.61738**
	K-medioides	0.51905
	DBSCAN	0.59444
No Recesión	K-means	0.55837
	K-medioides	**0.55995**
	DBSCAN	0.40994

Por lo anterior se puede decir que el algoritmo más adecuado para el análisis de clustering para la época de recesión es KMeans mientras que para la época de No recesión fue K-Medoids.

En el gráfico para Recesion se observan pares de clusters cercanos entre sí (0 -2) y (1 – 4), para el caso de la época de no recesión los clusters se pueden observar diferenciados, pero bastante cercanos entre grupos.



De acuerdo con la información presentada en la conformación de cada cluster de la música escuchada en la época de recesión, es posible determinar que el cluster 4 contiene las canciones con una media de popularidad mayor respecto a los demás (0.12), soportado con un coeficiente de variación robusto. Este clsuter está caracterizado por una época de alto de desempleo respecto a los demás (1.9), con un bajo nivel de ahorro (-2.44) y altos niveles de consumo (2.77), la música esta caracterizada por ser parte del género pop (0.49), HipHop (0.34) y rock (0.11). Además, se puede notar tienen la duración más baja respecto a otros clusters (0.18), un lenguaje explicito moderado (0.23), baja bailabilidad (-0.30), de energía relativamente alta (0.31), una muy baja instrumentalidad (0.084) y con una temática más negativa que la mayoría de los clústeres de 0.15 (solo superado por la negatividad del cluster 1). Por su parte, el cluster 0 está representado por las canciones menos populares para las épocas de recesión (-0.25). Estás se caracterizan por ser canciones más largas (0.26), más explícitas (0.33), un poco más bailables, aunque menos enérgicas, y dos veces más positivas (0.33).

Para el caso de las características musicales en clusters de años sin recesión económica, se puede evidenciar que el cluster 2 ostenta la agrupación de muestras con media de popularidad más alta (0.0883). Las canciones en este grupo tienen un alto índice de bailabilidad (0.318) y un tempo más alto respecto a otros clsuters (0.06), con una locuacidad alta (0.21), baja instrumentalidad (-0.11) y acústica de (0.26). Si bien existe una amplia variedad de géneros que hacen parte de este cluster, el pop lidera con 0.46, seguido por HipHop (0.46), rock (0.039) y música latina con (0.036). El cluster con la música menos

popular es el 4, cuya música se caracteriza por tener un contenido de bailabilidad significativamente bajo (0.088), alta instrumentalidad (0.12), baja acústica (0.04) y un tempo bajo (-0.013). Los principales géneros que le componen son el pop, hiphop y rock. Esto da a entender que, más allá de los géneros que están de moda de acuerdo con la época, la variabilidad se encuentra en las características musicales con las que se componen las canciones.

Los algoritmos usados se escogieron mediante la evaluación del Silhouette score, estos fueron adecuados ya que, con un bajo costo computacional, se obtuvo información valiosa respecto al objetivo del proyecto.

Se relacionaron las preferencias musicales con la realidad económica que representaron las variables seleccionadas sin embargo la principal limitación es la escogencia adecuada del número de clusters.

En nuestro caso no se tenía una idea inicial de los valores k, para minimizar esta limitación se utilizó cluster jerárquico y el método de codo. Sin embargo, se podría recomendar un método como Score de Calinski-Harabasz que han mostrado resultados más exactos en otros estudios (Neira, 2021).

Si bien los valores del coeficiente de Silhouette no son bajos tampoco son tan cercanos a 1, indicando una buena oportunidad de mejora para la separación de grupos.

K-means adicionalmente puede ser sensible a los valores atípicos que se encontraron, lo cual se puede ver reflejado en el tamaño del cluster 3 que solo tiene una observación. Se puede seguir explorando el hecho de reasignar el valor k inicial ya que la inicialización es crucial para este algoritmo.

Cluster	Tamaño-Recesion	Tamaño No-Recesion
0	33	399
1	20	326
2	76	327
3	1	322
4	61	375

El algoritmo K-means ha sido muy estudiado y adaptado en análisis, por lo cual es valioso su uso. El principal reto en el caso de estudio fue la presencia de los valores atípicos, aunque se realizó preselección de las variables y estandarización. Se podría explorar en futuros estudios el tratamiento de estos atípicos en las variables donde son altos por ejemplo instrumentalidad. Otra idea es explorar un algoritmo más robusto a los atípicos, en el caso de recesión DBSCAN, cuyo coeficiente de Silhouette (0.59444) no está muy alejado del valor obtenido para K-means (0.61738).

En cuanto a K-medoids este algoritmo generó grupos mas homogéneos y con características similares, que permitieron sacar conclusiones acerca de los mismos, esto podría vincularse a que este algoritmo es más robusto ante la presencia de datos atípicos. El presente proyecto abre puertas a seguir explorando en búsqueda de relaciones interesantes entre los momentos económicos y las preferencias musicales para que la industria musical pueda producir acorde a las posibles necesidades y preferencias de la población en un momento dado.

Conclusión.

Con el trabajo se lograron identificar grupos con características similares para ambas épocas recesión y no recesión.Se exploraron 3 algoritmos de agrupamiento y de acuerdo con el Silhouette Score se escogió el mas adecuado. Para los datos de recesión K-means predomina al dar grupos mas compactos y diferenciados, respecto a DBSCAN y K-medoids.

K Medoids, fue el más adecuado para los datos de los años de No recesión.

Los resultados para el Silhouette score no muestran diferencias tan altas entre algoritmos. El ajuste según el score en promedio para todos los algoritmos es de 0.5432 con una desviación estándar de apenas 0.07.

Los grupos obtenidos con K-means, sugieren la presencia de datos atípicos que pueden afectar el agrupamiento por lo cual se recomienda en un próximo trabajo explorar el tratamiento de estos y la implementación de DBSCAN que fue la segunda opción según el Silhouette score obtenido.La característica que más diferencian la popularidad de las canciones entre épocas es la bailabilidad, la cual caracteriza canciones más populares en época de no recesión y a las menos populares en recesión.

Los géneros Pop, Hip-Hop y Rock siempre están liderando en las preferencias, independientemente de la época, no obstante, en los años sin recesión, se resalta la música latina con una demanda muy similar a la del rock.

Bibliografía.

Liu, C., & Chao, Z. Supervised learning and unsupervised learning on music data with different genres. In Proceedings of the 2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom), 2021, pp. 7-12. DOI: 10.1109/DataCom53700.2021.00008.

Kim, K., Yun, W., & Kim, R. Clustering Music by Genres Using Supervised and Unsupervised Algorithms. 2015. Available at: https://www.semanticscholar.org/paper/Clustering-Music-by-Genres-Using-Supervised-and-Kim-Yun/0d93faaeef4055a71137eaa1ca7261cbd4dfe172.

Poonia, S., Verma, C., & Malik, N. Music Genre Classification using Machine Learning: A Comparative Study. Journal Name, 2022, 13, 15-21. Available at: https://www.researchgate.net/publication/362619781_Music_Genre_Classification_using_Machine_Learning_A_Comparative_Study.

Moreno, R. (2018, Jul 09). Qobuz no es como Spotify, un mero grifo de canciones". *Actualidad Economica,* , 27. https://ezproxy.uniandes.edu.co:8443/login?url=https://www.proquest.com/magazines/qobuz-no-es-como-spotify-un-mero-grifo-de/docview/2066200064/se-2

Plan, L. E. (1986). *Análisis multivariado : método de componentes principales*. OEA. Secretaría General.

Quiala Fonseca, W. (2023). Agrupamiento de datos desde un enfoque paralelo. *Revista cubana de ciencias informáticas*, *17*(4).

Neira Hernández, S. (2021). Alternativas en Clustering espectral. Universidad de los Andes. Disponible en: http://hdl.handle.net/1992/55107

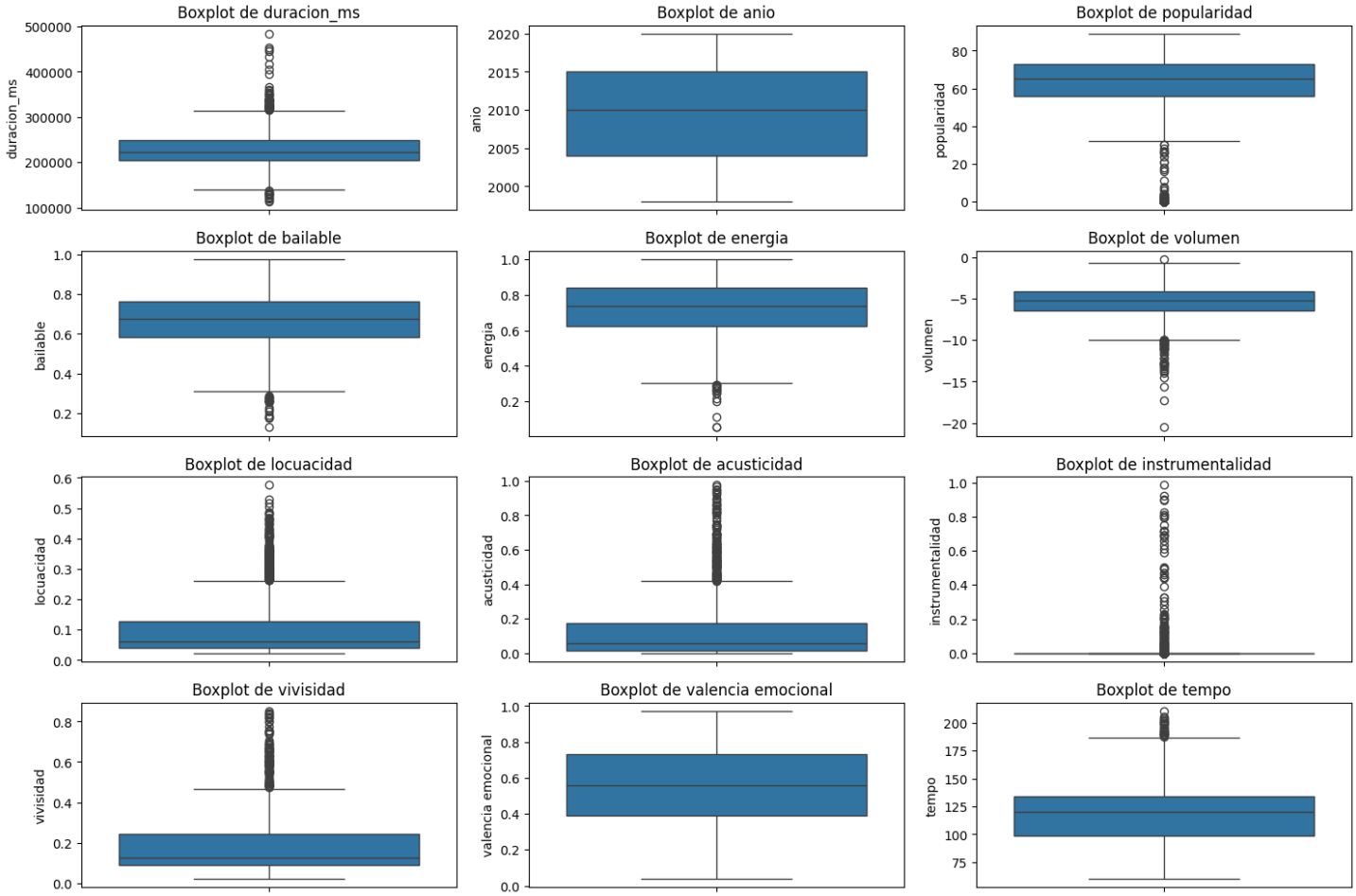
National Bureau of Economic Research. *US business cycle expansions and contractions*. https://www.nber.org/research/business-cycle-dating. Consultado el 9 de septiembre de 2024.

Anexo 1

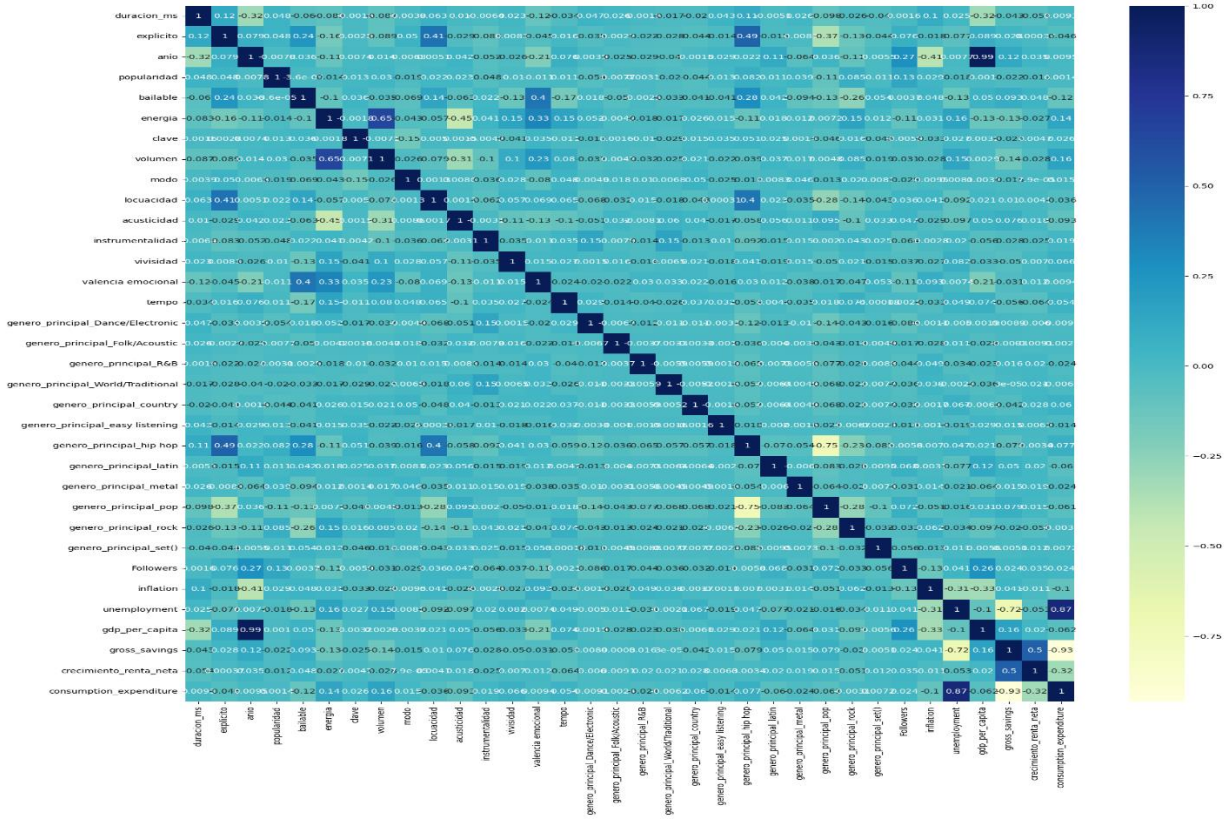
Descripción detallada de las variables.

Variable	Tipo de variable	Descripción
Artista	Categórica	Nombre del artista
Cancion	categórica	Nombre del track o nombre de la canción
Duracion_ms	numérica	Duración en milisegundos de la pista
Explicito	categórica	La letra o el contenido de una canción o un vídeo musical contienen criterios que podrían considerarse ofensivos o inadecuados para los niños.
Popularidad	numérica	cuanto mayor sea el valor, más popular será la canción.
Bailable	numérica	la bailabilidad describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general. Un valor de 0,0 es el menos bailable y 1,0 es el más bailable.
Energía	numérica	La energía es una medida de 0,0 a 1,0 y representa una medida perceptual de intensidad y actividad.
Clave (variable de codificación de las notas musicales)	Categórica	la clave en la que se encuentra la pista. Los números enteros se asignan a los tonos utilizando la notación estándar de clase de tono. P.ej. 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1. Corresponden a las notas musicales.
Volumen	numérica	el volumen general de una pista en decibeles (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar el volumen relativo de las pistas. El volumen es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db.
Modo	categórica	Modo indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. El mayor está representado por 1 y el menor es 0.
locuacidad	numérica	el habla detecta la presencia de palabras habladas en una pista. Los valores superiores a 0,66 = pistas que probablemente estén compuestas exclusivamente de palabras habladas. Los valores entre 0,33 y 0,66 = pistas que pueden contener música y voz, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores inferiores a 0,33 probablemente representen música y otras pistas que no sean de voz.
Acusticidad	numérica	una medida de confianza de 0,0 a 1,0 sobre si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.
Instrumentalidad	numérica	predice si una pista no contiene voces. Cuanto más cerca esté el valor de instrumentalidad de 1,0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 pretenden representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.
Vivisidad	numérica	Detecta la presencia de una audiencia en la grabación. Los valores de vivisidad más altos representan una mayor probabilidad de que la pista se haya interpretado en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.
Valencia emocional	numérica	Una medida de 0,0 a 1,0 que describe la positividad musical transmitida por una pista. Las pistas con valencia alta suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con valencia baja suenan más negativas (por ejemplo, tristes, deprimidas, enojadas).
Tempo	numérica	El tempo general estimado de una pista en pulsaciones por minuto (BPM).
Género_principal	categórica	Género musical de la pista.
Categoría_instrumentalidad	numérica	Es un campo calculado que implementa el campo instrumentalidad, si el valor es mayor o igual a 0.3 se cataloga como canción instrumental, en caso contrario es canción vocal
Followers	numérica	Numero de seguidores
día_semana	numerica	Día de la semana en que se realiza el lanzamiento de una canción
Inflation	numérica	La inflación mide el aumento sostenido en el nivel general de precios de bienes y servicios en una economía durante un período de tiempo. En términos simples, cuando la inflación sube, cada unidad de moneda compra menos bienes y servicios que antes. En Estados Unidos, la inflación se monitorea usando indicadores como el Índice de Precios al Consumidor (CPI) o el Índice de Precios al Productor (PPI). Un nivel moderado de inflación es normal en economías en crecimiento, pero niveles muy altos o bajos pueden indicar problemas económicos.
unenployment	numérica	La tasa de desempleo es el porcentaje de la fuerza laboral que está buscando activamente trabajo pero no puede encontrarlo. Se calcula dividiendo el número de personas desempleadas por la población activa (personas empleadas y aquellas que buscan empleo activamente). Es un indicador fundamental para medir la salud del mercado laboral de un país y su capacidad para generar empleo.
gross_savings	numérica	El ahorro bruto de un país se calcula restando el consumo de las personas y las empresas (gastos) de la renta disponible. Incluye: -Ahorros del gobierno. -Ahorros de las empresas (que retienen parte de sus ganancias en lugar de distribuirlas como dividendos). -Ahorros de los hogares (el dinero que no se gasta en consumo inmediato).
crecimiento_renta_neta	numérica	aumento o disminución del ingreso total disponible para los residentes del país después de considerar los pagos netos de ingresos al extranjero y la depreciación del capital. Es un indicador importante

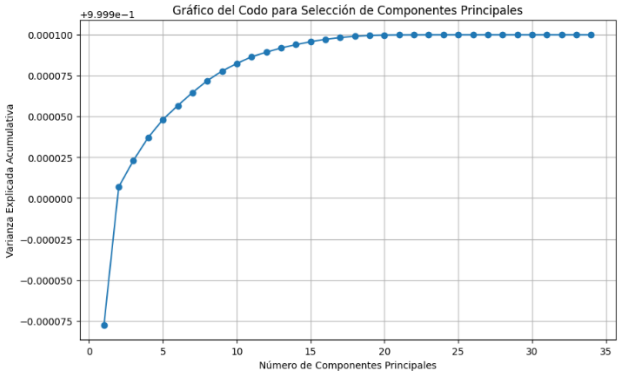
		que mide la salud económica de una nación, ya que refleja la cantidad de ingresos que efectivamente queda disponible para el consumo y la inversión interna.
consumption_expenditure	numérica	Este indicador representa el gasto total en bienes y servicios realizado por los hogares, el gobierno y las instituciones sin fines de lucro que sirven a los hogares (ISFLSH) dentro de una economía durante un período específico. En el caso de Estados Unidos, incluye: -Gasto de los hogares en bienes como alimentos, ropa, vehículos, y servicios como salud y educación. -Gasto del gobierno en bienes y servicios para el consumo colectivo (como defensa y seguridad). -Gasto de las ISFLSH, como iglesias, fundaciones y ONGs.
gdp_per_capita	numérica	El Producto Interno Bruto (PIB) per cápita es el valor total de todos los bienes y servicios producidos dentro de un país durante un año, dividido por la población total. Es una medida de la riqueza económica promedio de los individuos en una nación. En el contexto de Estados Unidos, es un indicador utilizado para comparar el nivel de vida promedio de los ciudadanos con el de otros países. Un PIB per cápita alto generalmente indica una economía más desarrollada y una mejor calidad de vida.

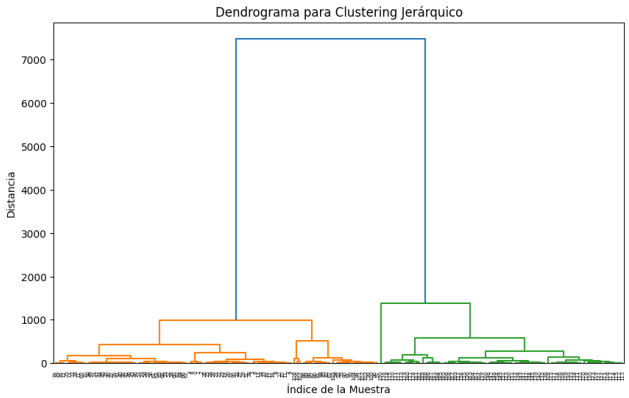


Anexo 2. Diagrama de correlación de las variables seleccionadas.



Recesión





No recesion

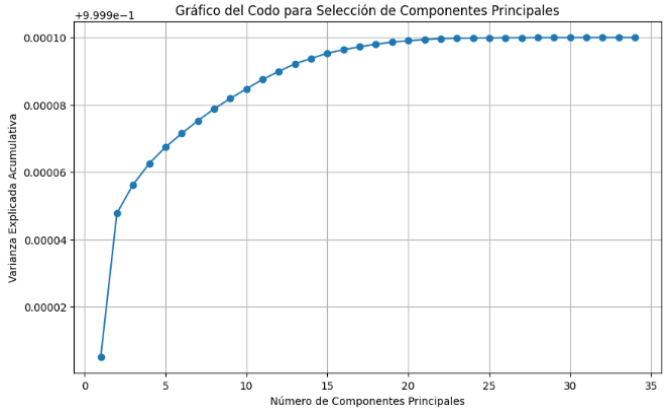
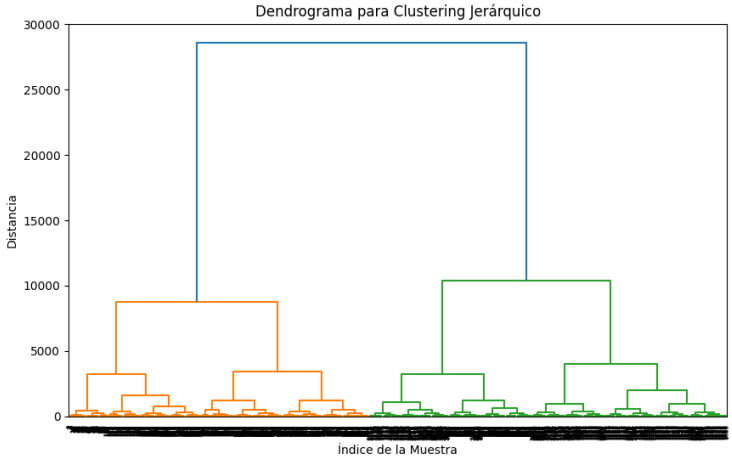


Gráfico del Codo para Selección de Componentes Principales
+9.999e−1
0.00010
0.00008
0.00006
0.00004
0.00002
Varianza Explicada Acumulativa
0
5
10
15
20
25
30
35
Número de Componentes Principales


Dendrograma para Clustering Jerárquico
30000
25000
20000
15000
10000
5000
0
Distancia
Índice de la Muestra